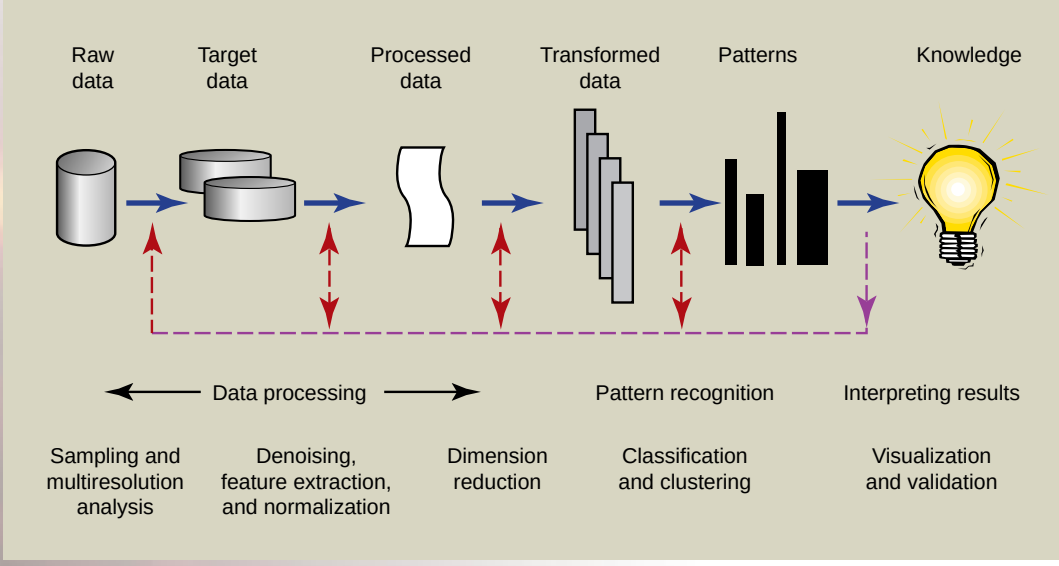
Mining Data for Gems of Information

MINING is an arduous, time-consuming business. Sometimes, tons of material must be excavated to uncover ounces of precious metals or gems. The computational equivalent of old-fashioned, down-in-the-dirt mining is data mining. Whether the search is for metals or information, the task is similar. In data mining, trillions of bytes of data must be sifted to find a handful of precious numbers or images.

As computers grow in speed, number-crunching capabilities, and memory, scientific researchers are edging into data overload as they try to find meaningful ways to interpret data sets holding more information than the U.S. Library of Congress. According to Livermore computer scientist Chandrika Kamath, "The problem has its roots in the many advances in technology that allow scientists to gather data from experiments, simulations, and observations in ever-increasing quantities," says Kamath. "In many scientific areas, the data sets are so enormous and complex that it is no longer practical for individual researchers to explore and analyze them by hand. When the sets get so large, useful information is easily overlooked, and the data cannot be fully utilized."

To address this problem, Kamath and a small team of Livermore researchers are developing Sapphire—a semiautomated, flexible data-mining software infrastructure. Sapphire shows great promise in helping scientific researchers plow through enormous data sets to turn up information that will help them better understand the world around us, from the makeup of the universe to atomic interactions. Sapphire is funded by the Laboratory Directed Research and Development program and the Department of Energy's Accelerated Strategic Computing Initiative (ASCI).

Data mining is not a new field. In the commercial world, it is used to detect credit card fraud and computer network intrusions; reveal consumer buying patterns; recognize faces, eyes, or fingerprints; and analyze optical characters. At Lawrence Livermore, the terascale computing environment created by ASCI as well as the prolific use of several different types of sensors have created great interest in large-scale, scientific data-mining efforts such as Sapphire. Kamath and her team envision that Sapphire will be applicable to a variety of scientific endeavors, including assuring the safety and



Sapphire, a data-mining infrastructure developed at Lawrence Livermore, is an iterative and interactive process designed to help scientific researchers uncover patterns in large data sets.

Preprocessing Techniques

Sampling: Selecting a subset of data items. Sampling is a widely accepted technique to reduce the size of the data set and make it easier to handle. However, in some cases, such as when looking for something that appears infrequently in the set, sampling may not be viable.

Multiresolution analysis: Another technique to reduce the size of the data set. With multiresolution analysis, data at a fine resolution can be "coarsened," which shrinks the data set by removing some of the detail. By preserving the detail, the transformation can be reversed.

Denoising: A technique that removes "noise" in images or data. It can be used to sharpen a fuzzy picture or aid in character recognition (differentiating a "6" from a "b" in text, for instance).

Feature extraction: A technique to extract relevant features from the raw data set. In credit card fraud, for instance, an important feature might be the location where a card is used. Thus, if a credit card is suddenly used in a country where it's never been used before, fraudulent use seems likely.

Dimension reduction: Reducing the number of features used to mine data, so only the features best at discriminating among the data items are retained.

Normalization: A technique used to level the playing field when looking at features that widely vary in size as a result of the units selected for representation.

reliability of the nation's nuclear weapons, nonproliferation and arms control, climate modeling, astrophysics, and the human genome effort.

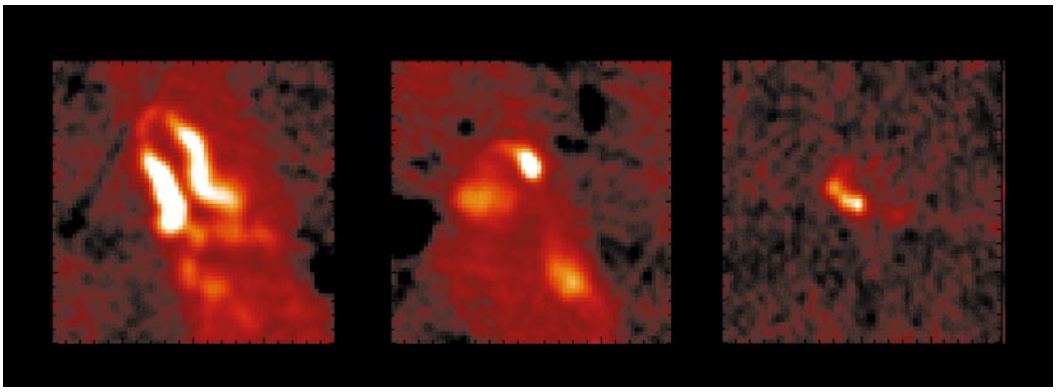
Data Mining Step by Step

Data mining starts with the raw data, which usually takes the form of simulation data, observed signals, or images. These data are preprocessed using various techniques such as sampling, multiresolution analysis, denoising, feature extraction, and normalization. (See the box at the left.)

Once the data are preprocessed or "transformed," pattern-recognition software is used to look for patterns. Patterns are defined as an ordering that contains some underlying structure. The results are processed back into a form—usually images or numbers—familiar to the scientific experts who then can examine and interpret the results.

To be truly useful, data-mining techniques must be scalable. "In other words," says Kamath, "when the problem increases in size, we don't want the mining time to increase proportionally. Making the end-to-end process scalable can be very challenging, because it's not just a matter of scaling each step but of scaling the process as a whole. For instance, the raw data set may be 100 terabytes, and as the data move through the data-mining process, the process decreases the data set size in ways we cannot predict. By the end of the process, we may have a resulting data set that's only a few megabytes in size."

To test and refine their algorithms, Sapphire researchers teamed up with Laboratory astrophysicists who were examining data from the FIRST (Faint Images of the Radio



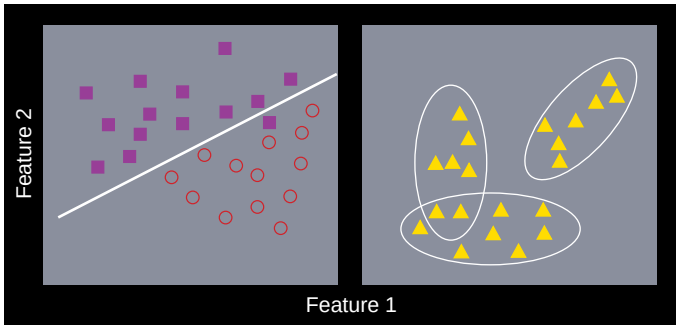
Radio-emitting galaxies with a bent-double morphology can appear completely different (as with the three shown in this figure), complicating the task of identification.

Sky at Twenty Centimeters) sky survey. This survey, which was conducted at the Very Large Array in New Mexico, seeks to locate a special type of quasar (radio-emitting stellar object) called bent doubles. The FIRST survey has generated more than 22,000 images of the sky to date. Each image is 7.1 megabytes, yielding more than 100 gigabytes of image data in the entire data set. Searching for bent doubles in this mountain of images is as daunting as searching for the needle in the proverbial haystack.

Mining Bent Doubles

The first step in applying data mining to this astrophysical search was to identify what features are unique to radio-emitting bent doubles. "Extracting the key features is essential before applying pattern recognition software," explains Kamath. "Although data exist at the pixel level (or at the grid level in mesh data), patterns usually appear at higher or coarser levels. The features—which can be any measurement—must be relevant to the problem, insensitive to small changes in the data, and invariant to scaling, rotation, and translation. Identifying the best features can be a time-intensive step, but it's a very important one."

Sapphire researchers worked with astrophysicists to draw up a list of features useful in identifying bent doubles. Such features included the number of "blobs," the spatial relationships of the blobs, and the peak intensity of the radio waves detected from each blob. "A parallel concern is to reduce the number of features to a relatively small set that will still provide accurate results," says Kamath. She notes that every additional feature used in pattern recognition on a terabyte data set adds enormously to the computational time and effort.



Classification vs clustering. In a simplistic example of classification (left), the algorithm tries to find the function (in this case, a line) that best separates the two classes. In clustering (right), the data are "plotted," and then clumps are identified to describe the data.

Once preprocessing is complete, the transformed data are input to pattern-recognition software. Two types of general pattern-recognition techniques used in data mining are classification and clustering. In classification, the algorithms "learn" a function that allows a researcher to map a data item into one of several predefined classes. In clustering, the algorithms work to identify a finite set of categories or clusters to describe the data. There are several different algorithms for classification and clustering, and frequently, both types of pattern recognition can be used within an application.

Once patterns are identified and translated by the Sapphire software back into a usable format, the results are examined by an expert. "We consider data mining to be a semiautomatic process because a human is involved in each step of the entire discovery process," explains Kamath. "The process is both iterative and interactive."

Gems Uncovered

Kamath and her team are pleased with how the data-mining algorithms tested out on the bent-double research—as are the astrophysicists. "Using our algorithms on the FIRST data, we identified a bent double previously overlooked by the astrophysicists in their manual search," said Kamath.

The data-mining algorithms in Sapphire are modular and easy to use in a variety of scientific applications and across diverse computer platforms. The beta release of this software to Lawrence Livermore users is scheduled for late 2000.

"We're also looking at what can be done to apply complex pattern recognition algorithms to data as they are being gathered," says Kamath. " For example, if one is looking for transient events—asteroids in astrophysics data or fraud in business transactions—the processing must keep up with the rate at which new data are acquired."

—Ann Parker

Key Words: Accelerated Strategic Computing Initiative (ASCI), data mining, Faint Images of the Radio Sky at Twenty Centimeters (FIRST), pattern recognition, Sapphire.

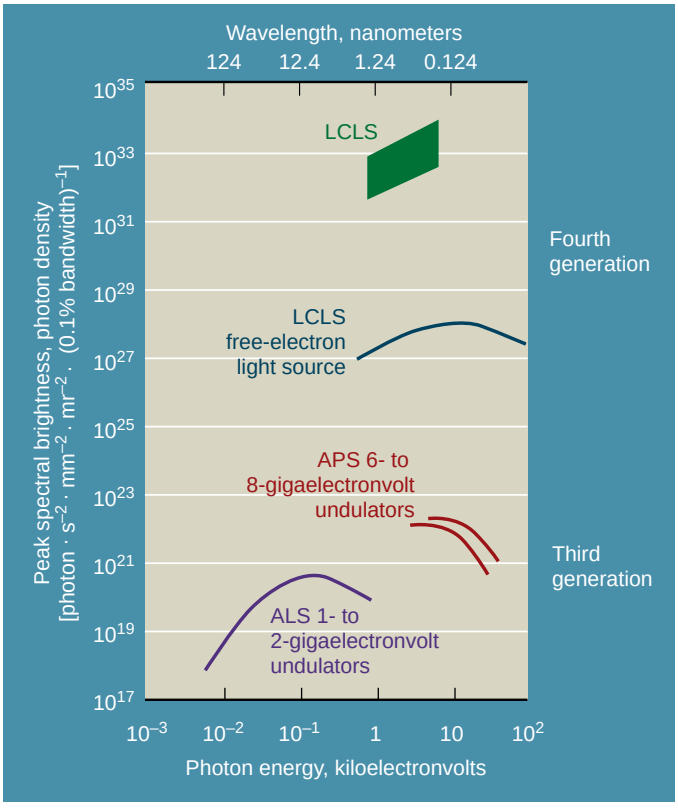
For more information contact Chandrika Kamath (925) 423-3768 (kamath2@llnl.gov).

More information about Sapphire can be found at www.llnl.gov/casc/sapphire/sapphire_home.html.

Low Emittance, High Brightness

A New X-Ray Light Source

X RAYS are handy for examining all kinds of materials, from our bones and lungs to high explosives and other ingredients in a nuclear weapon. If the x rays are intense enough and come in short enough pulses, they can supply information about the dynamic processes in many forms of condensed matter, such as solid materials, liquid crystals, and extremely dense plasmas. Using the Linac Coherent Light Source (LCLS)—an x-ray machine with unprecedented brilliance being considered for construction at the Stanford Linear Accelerator Center—researchers will be able to measure, for the first time, melting, recrystallization, and light-induced structural change on time scales down to a quadrillionth of a second.



Comparison of the brightness of the fourth-generation Linac Coherent Light Source (LCLS) with third-generation light sources—the Advanced Photon Source (APS) at Argonne National Laboratory and the Advanced Light Source (ALS) at Lawrence Berkeley National Laboratory.

The extraordinarily bright, short pulses of the LCLS have the potential to open new areas of science that are unimaginable given current scientific knowledge. The LCLS will make visible dynamic processes that can only be guessed at now. Upon completion in 2004, the new facility is certain to help solve problems in ultrahigh-energy-density physics, structural biology, fundamental quantum electrodynamics, warm dense matter, and high-field atomic physics, among others. The extreme brightness of the LCLS also means that results will be available much faster than before and will offer a level of detail that has been impossible to obtain with existing tools.

"For decades, we have studied nonlinear phenomena at optical wavelengths," says physicist Art Toor, who is leading the Livermore work on the LCLS. "But we've never had the tools to study nonlinear multiphoton processes in the x-ray region. That is tremendously exciting and opens the door to whole new regimes of research in physics, biology, and chemistry."

Protein crystallography, used to study the structure of proteins, is just one example of the research techniques that will benefit from the new x-ray source. Livermore and other biological research facilities use third-generation light sources to obtain images of molecules in a process that takes many hours of exposure time for each image. The shorter, brighter pulses of the LCLS will produce enough flux to image a molecule in a single pulse.

Livermore is part of the collaboration that is conducting research and development leading to this fourth generation light source. The LCLS is the next step beyond third-generation synchrotron radiation light sources, such as the Advanced Light Source at Lawrence Berkeley National Laboratory and the Advanced Photon Source at Argonne

National Laboratory. Third-generation light sources rely on storage rings where electrons traveling at nearly the speed of light are forced into a circular path by magnets. When the electrons pass through a magnetic structure called an undulator, they emit soft x rays that shine down beamlines to experimental stations.

The LCLS, in contrast, will use a linear accelerator rather than a circular one. It will also be home to the first x-ray free-electron laser, made possible by recent progress in undulator technology and in forming high-brightness, short-duration electron bunches in accelerators. The light from the LCLS will come in wavelengths smaller than the size of an atom. These hard x rays can be superior to longer-wavelength soft x rays for studying matter. The laser light will be fully coherent across the beam and 10 billion times brighter than the x-ray beams produced at the Advanced Light Source and its third-generation cousins. (Brightness is a measure of photon density, as shown in the figure on p. 23). The pulses will also be 100 times shorter than those of today's machines.

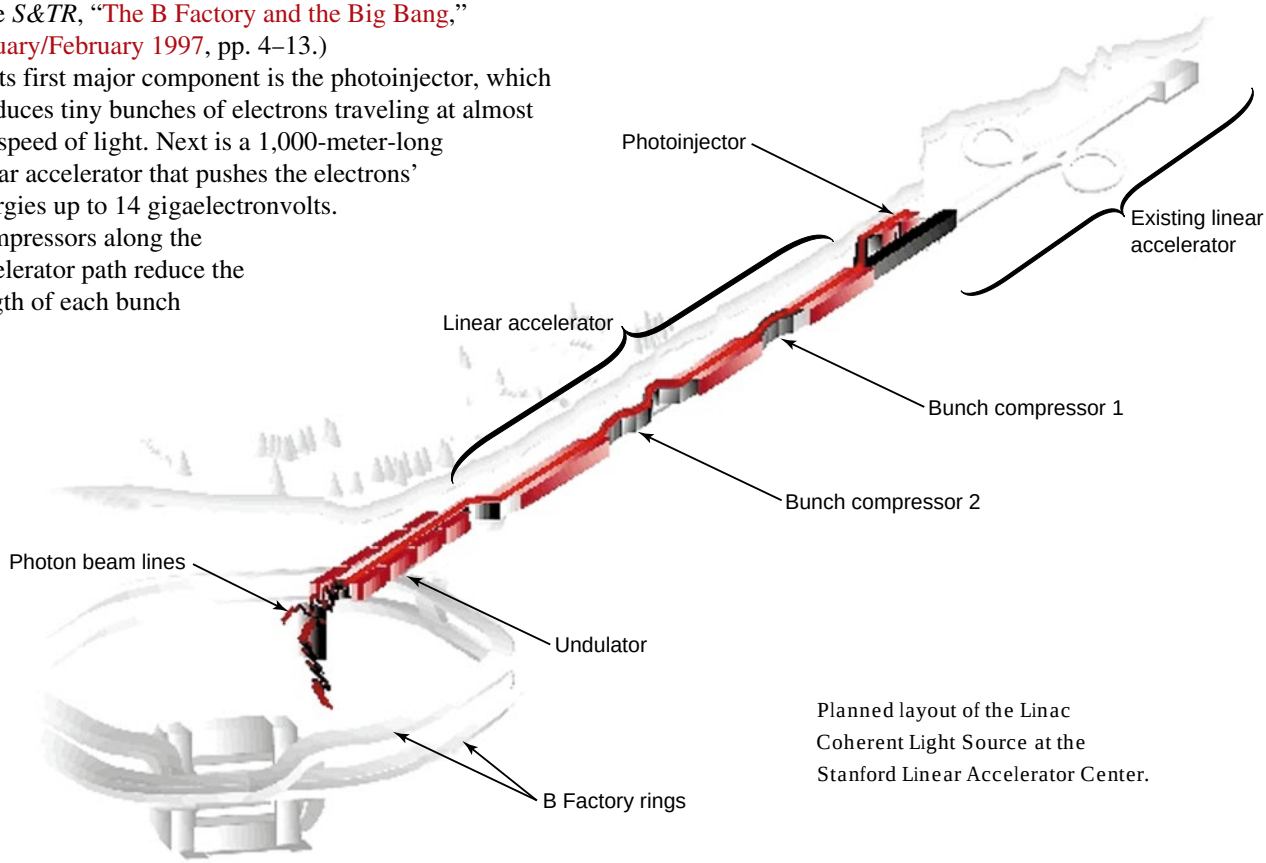
The Key to Success

The LCLS will be built around the portion of the Stanford Linear Accelerator that is not being used by the B Factory. (See *S&TR*, "The B Factory and the Big Bang," January/February 1997, pp. 4–13.)

Its first major component is the photoinjector, which produces tiny bunches of electrons traveling at almost the speed of light. Next is a 1,000-meter-long linear accelerator that pushes the electrons' energies up to 14 gigaelectronvolts. Compressors along the accelerator path reduce the length of each bunch by a factor of 30 to increase their peak current. Then the electrons enter an undulator, a vacuum chamber just 5 millimeters across and about 125 meters long that is lined with 7,000 magnets arranged in alternating poles. In this narrow channel, the magnetic fields push and pull on the electron bunches, causing them to emit x rays that in turn force the electrons into even tinier microbunches that release x-ray photons in a bright, coherent beam. Optical devices beyond the undulator manipulate the direction, size, energy, and duration of the x-ray beam and carry it to whatever experiment is under way.

Key to making this machine work is the low emittance of the electron beam injected into the accelerator. Emittance is a function of the diameter and divergence of a beam. A small beam with a wide spread has been easy to achieve, but a small beam with narrow spread has typically been difficult to produce. New photoinjector technology can produce a narrow, bright beam of electrons with emittance several times lower than previously achieved.

When the accelerated beam enters the undulator, interaction with the magnetic fields there causes x rays to appear. As the electron bunches move down the undulator,



Planned layout of the Linac Coherent Light Source at the Stanford Linear Accelerator Center.

the electron beam and the growing amount of x radiation interact more and more. More x rays produce more bunching, which produces more x rays, which makes the microbunches smaller and smaller, and so on. This chain reaction finally results in saturation of the x-ray beam to produce a narrowband, coherent beam of light, or laser, that is about 10 billion times brighter than the light from any other light source today. Also present is broadband spontaneous radiation about 10 thousand times brighter than that from any other light source.

Most other free-electron lasers store the light from many passes of the electron beam through the undulator in an optical cavity before putting it to use. The LCLS will require just a single pass by the electron beam through the undulator, thanks largely to the low emittance of the electron beam at the front end of the system.

Optics by Livermore

Livermore is part of a consortium with the Stanford Linear Accelerator Center, the University of California at Los Angeles, and Los Alamos, Brookhaven, and Argonne national laboratories that is developing the LCLS. Each institution is responsible for a different part of the overall project.

In one project, Livermore scientists are working with colleagues at Brookhaven National Laboratory to demonstrate the new technology for LCLS and produce the first free-electron laser in the visible wavelength. The Livermore team designed the 4-meter undulator, vacuum system, and other portions of the project.

For the LCLS at Stanford, Livermore will design and fabricate the x-ray optics downstream of the undulator. Says Toor, "The high peak power, full transverse coherence, and very short pulse lengths combine to make the optics for the LCLS a real challenge. We also are designing optical systems to accommodate a variety of experiments. Some require submicrometer focus at very high intensity, others require only coherence, and still others require illuminating large areas at much lower light levels."

A critical element in the optical system that Livermore scientists are working on is the absorption cell, which intercepts the beam after it leaves the undulator. The cell attenuates the beam's power to levels manageable with conventional optics and provides a transition to power densities that match the needs of the various experiments. The cell can also completely remove the free-electron laser light for experiments that use only the spontaneous radiation. Both the spontaneous and coherent x radiation pass through an ultrahigh-vacuum system to the experimental areas, which may ultimately be as much as a kilometer away. Shielding will protect personnel and experiments from bremsstrahlung, the gamma radiation that results from high-energy electrons interacting with matter.

This work is opening new territory. Virtually no information exists now on the interaction of extremely high levels of hard x rays with matter. If the Department of Energy approves construction of the LCLS, the beginning of testing and experimentation in 2004 will herald a brave new world in physics.

—Katie Walter

Key Words: free-electron laser, Linac Coherent Light Source (LCLS), linear accelerator, Stanford Linear Accelerator Center.

For further information contact
Art Toor (925) 422-0953 (toor1@llnl.gov).